Journal of College Access

Volume 5 | Issue 1 Article 4

1-2020

Will I Get In? Using Predictive Analytics to Develop Student-Facing Tools to Estimate University Admissions Decisions

Matt S. Giani
University of Texas at Austin, matt.giani@austin.utexas.edu

David Walling
University of Texas at Austin, walling@tacc.utexas.edu

Will I Get In? Using Predictive Analytics to Develop Student-Facing Tools to Estimate University Admissions Decisions

Authored by
Matt S. Giani (*University of Texas at Austin*)
David Walling (*University of Texas at Austin*)

ABSTRACT

A sizeable number of low-income high school graduates enroll in colleges less selective than their academic qualifications would allow or forgo postsecondary altogether despite being college-ready. One potential cause of this "undermatching" is that some students have limited access to information about their college options. We hypothesize that providing students with more and better information about the relationship between their academic preparation and college options may promote college-going. The purpose of this study was to develop a predictive model of admissions to public 4-year institutions using data from Texas' statewide longitudinal data system in order to build a student-facing tool that predicts admissions decisions. We sought to include only variables for which students have some control over, namely academic characteristics, but compared the predictive accuracy of this reduced model to more complex models that include demographic variables commonly used in higher education research. We show the reduced model successfully predicts admissions decisions for approximately 85% of applications. The addition of demographic variables, despite showing a statistically significant better fit of the data, do not substantively change the predictive accuracy of the model. We include a demonstration of a data visualization tool built on this predictive model using the open-source *R* statistical software that can be used by students, parents, and educators. We also discuss causes for both optimism and caution when using predictive modeling to develop student-facing tools.

Keywords: admissions, predictive modeling, student-facing

The United States has made considerable progress in increasing college access rates for all racial and socioeconomic subgroups since the Civil Rights era (National Center for Education Statistics, 2016), but there is evidence that baccalaureate completion rates have actually declined over time (Bound, Lovenheim, & Turner, 2010) and disparities in baccalaureate attainment have remained stubbornly persistent (NCES, 2016). A common explanation of this phenomenon is that many high school graduates may not be academically prepared to access and succeed in college, and disparities in "college readiness" may contribute to inequitable attainment outcomes (Adelman, 1999, 2006; Cabrera & La Nasa, 2000, 2001; McPherson & Shapiro, 1998; Terenzini, Cabrera & Bernal, 2001). However, a growing body of literature has also identified the issue of "undermatch," in which students enroll in postsecondary institutions that are less selective than those for which they are qualified or forgo postsecondary enrollment altogether (Bowen, Chingos, & McPherson, 2009; Roderick et al., 2008; Hoxby & Avery, 2012; Roderick, Coca, & Nagaoka, 2009; Smith, Pender, & Howell, 2012).

Researchers continue to investigate the diverse causes of undermatch, but a compelling hypothesis is that students with limited access to information about their college options may be more likely to undermatch. Indeed, studies have found that high-achieving students are less likely to apply to and enroll in selective colleges if they attend small or rural high schools with fewer high achievers (Hoxby & Avery, 2012), and interventions that provide high-ability students with greater information about institutions which they are qualified for have been found to significantly increase the likelihood that they apply to selective colleges, are admitted, and matriculate (Hoxby & Turner, 2013). These studies suggest that providing students with more accurate information about their college options may be an effective strategy for increasing college-going overall and decreasing equity gaps in college access. Towards this end, the purpose of this study is to use predictive modeling to develop a student-facing tool designed to estimate the likelihood of university admission using data from Texas' longitudinal student data system. The goal was to include only variables for which the student has some control over, namely variables tied to their academic achievements. These include variables such as GPA, SAT/ACT scores, the high school graduation plan a student completes, and the number of advanced and dual-credit courses passed in high school. We explicitly desire to exclude variables for which the student does not have control, such as race, sex, and parents' socioeconomic status. However, such variables are commonly used in higher education research. Thus, in order to justify their exclusion from our modeling approach, we must first verify that their influence does not greatly affect the predictive performance of our reduced model. We show this by comparing a full model combining the desired academic variables and the control variables to a reduced model containing only the variables of interest. We demonstrate that the reduced model performs as well as the full model and correctly predicts admissions decisions for roughly 85% of public university applications in Texas.

Academic Resources, Information, and Undermatch

There is broad consensus in the literature that academic resources influence college access and completion rates, readiness for college is unequally distributed across racial/ethnic and SES groups, and disparities in academic preparation at least partially explain inequities in baccalaureate attainment (Adelman, 1999, 2006; Cabrera & La Nasa, 2000, 2001; Kim, 2004; McPherson & Shapiro, 1998; Terenzini, Cabrera & Bernal, 2001). However, a growing body of research has highlighted the magnitude and significance of "undermatch," or the phenomenon in which students enroll in postsecondary institutions significantly less selective than those for which they are qualified or forgo postsecondary enrollment altogether despite being college-ready (Bowen, Chingos, & McPherson, 2009; Roderick et al., 2008; Hoxby & Avery, 2012; Roderick, Coca, & Nagaoka,

2009; Smith, Pender, & Howell, 2012). Although some studies suggest there may be risks of "overmatching" given that students who overmatch may be surrounded by peers with greater academic qualifications than them (Sander & Taylor, 2012; Thernstrom & Thernstrom, 1997), the majority of studies in this vein have concluded that overmatching increasing the likelihood of attainment (Alon & Tienda, 2005) or, conversely, under-matching decreases the odds of attainment (Bowen, Chingos, & McPherson, 2009).

Studies have shown that low-income students are significantly less likely to apply to a four-year institution compared to their high-income peers, even when controlling for academic readiness (Cabrera & La Nasa, 2001; Author, 2015; Hurtado, Inkelas, Briggs, & Rhee, 1997; Pallais & Turner, 2006). In their analysis of students' pre-college pathways using data from NCES' National Education Longitudinal Study of 1988, Cabrera and La Nasa (2001) identified the rates at which students from different SES backgrounds became "college-qualified," graduated from high school, and applied to postsecondary institutions. Out of the pool of college-qualified high school graduates, the authors noted that only 65.5% of student from the lowest-SES background applied to a four-year institution, 16% below the national rate for college-qualified students and 22% below the rate for college-qualified students from the highest-SES background. In other words, only two out of three college-qualified low-SES graduates applied to a four-year institution, compared to nearly nine out of ten high-SES graduates who were college-qualified. However, the authors concluded that the chances of lowest-SES students enrolling in a four-year institution "improve dramatically to the point of closely resembling the national average and the rate for highest-SES students" once low-SES students complete the task of submitting an application to a four-year college or university (p. 121).

Hoxby and Avery (2012) reached similar conclusions when analyzing the rates at which very high achievers, or students with an SAT score in the top ten percent of the national distribution and who had at least a 3.5 GPA in high school, applied to selective colleges. The authors found that "a large number--probably the vast majority--of very high achieving students from low-income families do not apply to a selective college or university" (p. 1). However, these low-income high-achievers exhibited different application patterns. The group of high-achieving low-income students the authors defined as "income-typical" had low application rates and rarely applied to selective institutions, while "achievement-typical" students applied to more colleges and more selective colleges, mirroring the application patterns for high-income high-achievers. Put differently, very few high-achieving low-income students apply to a broad range of schools, many of which are selective, which is the common application behavior for high-income high-achievers. The authors also found that income -typical students were more likely to attend high schools with few other high achievers and which had a weak history of graduates

attending selective colleges. In other words, despite being high-achieving, these students were less likely to have the information and support need to promote their college aspirations and application behavior. Subsequent interventions designed to identify these high-achieving low-income students and provide them with greater information about their college options have been found to significantly increase these students college application rates, rates of application to selective institutions, and the total number of applications they submitted (Hoxby & Turner, 2013). More importantly, high-achieving low-income students have been found to the admitted to selective colleges at rates roughly equivalent to their high-income peers (Hoxby & Avery, 2012), and these interventions did in fact increase the selectivity of institution that low-income students matriculated to (Hoxby & Turner, 2013). These findings suggest providing greater information to high-achieving low-income students about their college options may not only promote their college application rates and the selectivity of colleges to which they apply, but may also promote their college enrollment, decrease undermatching, and potentially reduce inequities stemming from socioeconomic background in the selectivity of colleges students enroll in.

Although this line of research is promising, the proportion of low-income students that fall into the high-achievement category as defined by Hoxby and Avery (2012) is quite small – they estimate between 25,000 and 35,000 students in each national cohort of high school graduates fall into this category. While encouraging high-achieving low-income subgroup's college aspirations and applications is important, focusing exclusively on students with the academic qualifications needed to gain access to the most selective schools in the country may be an overly narrow approach. However, it is also much easier to design interventions like the one piloted by Hoxby and Turner (2013) for a few thousands students rather than the millions who graduate high school each year. This problem motivated the current study. We sought to develop a tool to accurately estimate students' likelihood of college admission that could be used by educators, students, and students' families to make more informed decisions about applying to college. Our goal was to make this tool useful to all students, not just very high achievers. And we also believed more generally that providing students with better information about the relationship between their academic performance and their likelihood of admittance into specific colleges and universities might motivate students to pursue a more challenging high school curriculum, earn better grades, and the like. However, this tool would only be useful if it was a valid and reliable predictor of students' admissions decisions. The sections to follow describe our methodological approach for building and validating the underlying statistical models which the tool is founded upon.

Methods

Data Source and Access

The data used in this study was made available by the Texas Education Research Center (ERC) at The University of Texas at Austin. The ERC houses several datasets collected from the Texas Education Agency (TEA), Texas Higher Education Coordinating Board (THECB), and Texas Workforce Commission (TWC) and makes it securely available for scientific inquiry and policy making purposes. Access to the data can be acquired by submitting a research proposal to the ERC Joint Advisory Board, which reviews proposals based on whether data needed to address the research questions is available in the ERC, the strength of the proposed methods, and the potential benefits of the research to the state of Texas. Access to the data can also be granted directly by the Texas Legislature, as is the case for the current study.

Data collected by THECB through the ApplyTexas application system was used to document students' applications and admissions decisions. All public universities in the state are required to use ApplyTexas to accept applications from Texas high school graduates (see applytexas.org). Community colleges also use ApplyTexas but are not required to report data on applications to the state, preventing us from analyzing applications to community colleges. This dataset contains a record for every application students submitted through ApplyTexas, the admissions decision of the institution, and a host of other background demographic and academic variables. Specifically, data on high school ranking and SAT/ACT scores (discussed below) are collected through this system. It is important to note that in addition to ApplyTexas, institutions may offer additional application systems, such as the Common Application or institution-specific admissions processes, and students who apply to universities through those systems are not recorded in the ApplyTexas dataset. However, anecdotal evidence suggests that the vast majority of Texas high school graduates who apply to Texas public universities use ApplyTexas.

Our cohort was defined using the Texas Education Agency's (TEA) high school graduation data. This dataset includes a record for every student who completed high school during a particular year. Data on students' high school transcripts was collected by TEA. This data source includes information on the title of each course students attempted in high school, whether the course was advanced, whether the course was dual-credit, the subject of the course, whether the student passed the course, and the number of credits the student earned from the course. One idiosyncrasy of the dataset is that numerical course grade information was collected and reported during the 2010-11 and 2011-12 school years but for no other years before or after. Given our use of a 2014 cohort of high school graduates (sample described below), we had data on grades for students' freshmen and sophomore years of high school but not their junior or senior years.

Additionally, the dataset only contains information for courses taken through Texas high schools, so students who transferred into Texas during high school would not have their prior course taking recorded in the data. The TEA data also contains a file with detailed information on students' demographic backgrounds. This dataset was used to determine students' race/ethnicity, sex, and economic background (free-or-reduced lunch eligibility). Although the ApplyTexas dataset also contains information on students' socio demographic backgrounds, certain variables appear to have significant amounts of missing data whereas the TEA data was far more complete.

Sample

The sample used in the current study is a cohort of students who graduated from a Texas high school in 2014 and who applied to at least one public university in the state of Texas for admissions during the fall 2014 semester. Of the 302,269 students in the graduating class, 103,860 students (34.36%) submitted at least one application, and 200,973 individual applications were submitted. Demographically, the sample was 6.1% Asian, 15.0% Black, 41.7% Hispanic, 33.0% White, and 4.1% other (Native American/Alaskan Native, Native Hawaiian/ Pacific Islander, and multiracial students were combined into this category due to their small sample sizes), 55.1% female compared to 44.9% male, and 41.3% economically disadvantaged compared to 58.7% non-disadvantaged.

From the original sample of applications, we excluded all instances where the students withdrew their applications since they did not receive an admissions decision in that case, as well as applications where the student was admitted under the top ten percent policy. This was done for two reasons. First, all students in the top ten percent receive automatic admission, meaning there is no variation in the outcome variable for this subgroup. A predictor variable representing whether students were in the top ten percent would therefore be dropped from the statistical model. Second, because these students are guaranteed admission, the tool we developed would be irrelevant to this population. Excluding top ten percent students, withdrawn applications, and a small percentage of students with missing data (discussed below) left 110,620 application records. We further split this sample into training and test sets at a ratio of 80/20, with the test set used to analyze the performance of the models developed on the training set.

Variables

The outcome variable in the study is whether students were admitted to a public university in Texas to which they applied. The university applications dataset includes a variable that indicates the admissions decision for each application. This variable has seven possible values: 1) accepted and ranked in the top 10% of graduating class; 2) accepted and ranked in the 11-25% of graduating class; 3) accepted on provisional basis, met requirements; 4) accepted on provisional basis, did not meet requirements; 5) accepted based on other

criteria; 6) rejected; 7) student withdrew application. As mentioned above, students accepted through the top ten percent rule and those with withdrawn applications were excluded. The original admissions decision variable was converted into a dichotomous variable, with the rejected (6) value being recoded into not admitted ("0") and the values of 2-5 being recoded as admitted ("1"). The five academic variables of interest included in the models are the student's high school GPA, ACT/SAT score, number of advanced courses, number of dual credit courses, and high school graduation level. Of the primary variables, GPA was the only one to present particular difficulties. As mentioned above, grade data was only available for the years 2011 and 2012, years when our cohort would have been freshman and sophomores, and it was from these values that GPA was calculated. Because some of the cohort were not attending a Texas school during these years, GPA was missing for those application records (n = 2,766, or 2.4% of sample) and were dropped from the analysis.

SAT/ACT scores were recorded in the ApplyTexas application. Some students only reported an ACT score, some reported an SAT score, some reported both, and some reported neither (14,621 application records, or 7.3% of the total sample of 200,973 applications). In order to include a single variable in the model, SAT scores were converted to the ACT score range of 11-36 using SAT-ACT concordance tables (College Board, 2016). It is also noted that multiple applications from the same student may contain different values for this variable, indicating the student retook the given test and submitted improved scores. We used the SAT/ACT score the student submitted to the institution she or he applied to, rather than the highest score they submitted across institutions. Applications without SAT/ACT scores were dropped from the sample.

The number of advanced and dual-credit course variables are measured by counting the number of credits students earned for courses indicated as advanced or dual-credit in the TEA data. A full year course is generally worth one credit in the data but may be broken up into two semester-long courses each worth 0.5 credits, for example. Although schools and districts may have used different criteria for determining whether students passed courses, failed courses were awarded zero credits and were therefore excluded in the calculation of these variables. At the time when this cohort was graduating from high school students could earn one of four different types of high school diplomas: distinguished, recommended, minimum, and individualized education plan (IEP). Roughly 70% of the cohort completed the recommended plan. The distinguished plan included additional rigorous courses and approximately 15% of students earned that diploma. The remaining 15% of students completed the minimum plan or an IEP. Most frequently, students with disabilities complete IEPs. Because of the small number of students earning IEPs, the minimum and IEP categories were collapsed into a single

category. This three-level variable (distinguished, recommended, and minimum/IEP) representing the diploma students earned was included in the models as an additional measure of curricular rigor. There were 35 public universities represented in the original dataset, but four of these were small schools where only a handful of applications were received from our cohort of students. We grouped all schools with < 100 applicants into an 'other' category. The statistical models include institutional fixed effects, which essentially use the institution's overall admission rate to adjust the students' baseline odds of admission.

As our primary purpose was to develop a student facing tool to estimate admissions decisions we desired not to include demographic variables in the models, both because students have no control over their demographic backgrounds and because we would not want students to see their odds of admission change depending on their race, SES, or sex. However, because prior literature has shown students' demographic characteristics at times shape their college-going behavior, we sought to further validate the tool by fitting statistical models that controlled for race, SES, and sex. Race has been grouped into 5 categories: White, Black, Asian, Hispanic and other (American Indian/ Alaskan Native, Native Hawaiian/Pacific Islander, Multiracial, or Unknown). Socioeconomic status was proxied with a binary variable indicating whether students qualified for free-or-reduced price lunch in high school. A dummy variable for males was included with the reference group being females, as Texas does not allow students to report non-binary gender identities.

Model Validity and Comparisons

In much quantitative educational research, and in particular studies that use some form of regression modeling, the primary interest is often the relationship between independent variables of theoretical import and the outcome. These relationships are assessed through the magnitude and direction of the coefficients, as well as whether the estimates are statistically significantly different than zero at whatever threshold the researcher chooses, most commonly $p < .05$. At times researchers present values such as R2, the proportion of variance in the outcome explained by the model, or various fit indices to assess how well the model fits the data, but rarely are those statistics the main focus of the research. However, in our case the accuracy and reliability of the model(s) are far more important than the relationship between individual predictors and the outcomes, given our goal of creating a tool students can reasonably rely upon to estimate admissions decisions. We therefore employed a variety of statistical techniques for assessing the validity and performance of these models.

We first checked for potential issues of multicollinearity, or when predictor variables in the model are highly related to each other (Belsley, Kuh, & Welsch, 1980; Greene, 2011), by computing variance inflation factors (VIF) for each of our models. The VIF values represent how much the variance is increased due to issues of multicollinearity. VIF values greater than 10 suggest the possibility that multicollinearity may be affecting the results,

although some statisticians argue that VIF values as high as 40 can still be tolerated without biasing the results (O'Brien, 2007). Nevertheless, all variables had a VIF value less than 10 for all models included in the study, suggesting limited threat of multicollinearity.

We then examined measures of accuracy of the models defined by their Receiver Operating Characteristics (ROC), including their sensitivity and specificity, as well as the related Area Under the Curve (AUC) values (Hosmer, Lemeshow, & Sturdivant, 2013). The ROC measures the accuracy of the model by classifying predictions based on whether they are above and below 0.5 and then comparing the predicted values to the actual outcome. For example, if a student has a 0.75 (75%) predicted likelihood of being admitted to a college but they were not admitted, that prediction would be considered inaccurate. The ROC summarizes the overall accuracy of the model, and predictions can be further classified based on the ability to detect true positives (sensitivity) and true negatives (specificity). The AUC essentially compares the models to one that would randomly classify cases. An AUC value of .5 means the model is no better than chance at predicting the outcome, while an AUC between .9-1.0 suggests excellent fit.
In addition to examining overall accuracy, we use a common metric known as the Brier score to explore other performance characteristics of the models (Brier, 1950; Murphy, 1973). In particular, we are interested in how well the models are calibrated, or how accurate they are over the entire range of values, i.e. the probability threshold for labeling a prediction for a student as 'accepted'. Two competing models could correctly predict the same number of events overall, but one may over predict events with high probability and correspondingly under predict those with low probability while the other is more accurate over the entire range of values. Brier scores range from 0.0-1.0, with values closer to 0.0 representing better calibration.

With ordinary least squares regression (OLS), the most common measure of model performance is R2, a value representing the amount of observed variability in the outcome explained by the given model. A directly analogous measure of model performance is not possible with logistic regression because the maximum likelihood calculation for logistic regression is not minimizing variance. In lieu of R2, a variety of 'pseudo' R2 values have been developed to provide similar metrics for logistic regression, with several producing R2 like values ranging from 0 to 1, but with slightly different interpretations. While there is no consensus on the best version of pseudo R2 values to use, one of the most common is the adjusted McFadden's pseudo R2, where values of this metric between 0.2-0.4 indicate excellent fit, and roughly correspond to values of 0.7-0.9 of the OLS version of R2 (McFadden, 1974).

Results

The models discussed in the results section are numbered as follows:

Model 1: Reduced
Model 2: Reduced+Sex
Model 3: Reduced+Race
Model 4: Reduced+Econ
Model 5: Reduced+Sex+Race
Model 6: Reduced+Sex+Econ
Model 7: Reduced+Race+Econ
Model 8: Reduced+Sex+Race+Econ

Each model includes all of the primary variables of interest, and differ only in which of the control variables they contain. Lower numbered models are said to be nested within higher models where the higher model contains all the variables of the nested model in addition to others. For instance, model 4 is nested in model 6, but not in model 5 as model 5 does not contain the economically disadvantaged variable. It was our desire to examine each of the possible combinations of the control variables, so automated variable selection such as step-wise methods where not utilized.

Model Summaries

Summaries of each model are provided in Table 1 on page 27. Apart from the predictor variables included in the table, the models also include university fixed effects which are not shown for conciseness. The primary academic variables of interest are statistically significant for each of the models under consideration. Although both advanced and dual-credit courses were found to be positively related to acceptance, of note is that advanced courses had roughly twice the benefit in terms of admissions compared to dual-credit courses.

The demographic control variables were also found to be statistically significant in every model in which they were included. Males had lower odds of admission compared to females, all racial/ethnic groups had lower odds of admissions compared to the reference category of Hispanics (although the coefficient for Whites was not statistically significant in the fullest models), and economically disadvantaged students were less likely to be admitted compared to non-disadvantaged students. We note that the addition of control variables had little effect on the estimated coefficients of the primary variables of interest.

Whereas the results showed that the variables included in the models were significantly related to students' odds of admission, of greater importance is the validity of the models. Table 2 presents the ROC statistics, including the overall accuracy of the models as well as their sensitivity and specificity, calculated using 10-fold cross validation. The results show that the models correctly classify roughly 84.0% of students overall, although the models are better at classifying true positives (91.1-91.2%) than true negatives (69.2-69.6%). Put differently, roughly 9.0% of students who did get into the institution they applied to would have been told that they would not get in (the false negative rate),

while roughly 30.0% of students who were not admitted would have been told that they would be (the false positive rate). Importantly, neither the overall accuracy of the models or their sensitivity and specificity vary appreciably regardless of the demographic variables controlled for, suggesting the model would be just as valid excluding demographic characteristics.

Table 2.
ROC Tests

	Accuracy	CI Lower	CI Upper	Sensitivity	Specificity
1	0.8375	0.8332	0.8418	0.9107	0.6917
2	0.8378	0.8334	0.8420	0.9112	0.6915
3	0.8386	0.8343	0.8429	0.9115	0.6935
4	0.8383	0.8340	0.8425	0.9117	0.6921
5	0.8383	0.8339	0.8425	0.9113	0.6927
6	0.8388	0.8345	0.8431	0.9116	0.6939
7	0.8384	0.8341	0.8427	0.9116	0.6926
8	0.8399	0.8356	0.8441	0.9123	0.6956

Table 3 provides the adjusted McFadden pseudo-R2, AUC, and Brier score values for all eight models. All three statistics provide strong support for the models' validity. The high pseudo-R2 and AUC values suggest strong accuracy of the models, and the relatively low Brier scores suggest the models are reasonably well calibrated across the range of predicted values. Again we see that the difference in performance for the most complex model versus the most parsimonious one is practically negligible, even when the demographic variables added to the models may be statistically significant given the large sample size.

Table 3.
Predictive Admissions Tool

	Adj.McFadden	AUC	Brier Scores
1	0.5097	0.9339	0.1920
2	0.5102	0.9340	0.1922
3	0.5106	0.9341	0.1919
4	0.5101	0.9341	0.1913
5	0.5111	0.9342	0.1916
6	0.5106	0.9341	0.1917
7	0.5108	0.9342	0.1914
8	0.5113	0.9342	0.1912

Finally, we have developed a prototype of an interactive visualization tool driven by our model. This tool allows students and counselors to explore the impact of academic performance and high school course-taking decisions on admissions rates to their selected colleges. A screenshot of this prototype is provided in Figure 1 on page 28. A web-link to this tool and the code used to generate the tool from the underlying statistical model is available upon request to the corresponding author.

Discussion

Low-income and URM students have been found to be significantly more likely than their high-SES peers to apply to and enroll in colleges that are significantly less selective than those for which they are qualified or forgo postsecondary altogether (Bowen, Chingos, & McPherson, 2009; Roderick et al., 2008; Giani 2015; Hoxby & Avery, 2012; Roderick, Coca, & Nagaoka, 2009). This is despite the fact that the vast majority of high school students aspire to attend a 4-year college, regardless of socioeconomic background and race/ethnicity (Author's

calculations using NCES' Datalab). Preliminary interventions providing high-achieving, low-income students with more and better information about the types of universities they are likely to gain admission to and the cost of attendance of these institutions have shown promising in increasing college application rates and reducing undermatch (Hoxby & Turner, 2014).

There are surely diverse causes of undermatch, but a compelling explanation is that students who undermatch may have limited information about their college options. Hoxby and Avery's (2012) analysis showed that what distinguished high-achieving, low-income students' college application patterns was the types of high schools they attended. "Achievement typical" students were more likely to attend high schools with other high achievers and where previous graduating cohorts had attended selective colleges, while "income typical" students were relatively isolated from other high achievers and attended high schools without a strong history of sending students to selective institutions. It is possible, then, that these students have insufficient knowledge about the types of institutions for which they are qualified, despite being academically prepared to succeed in college. Our goal in this paper was to develop a tool that estimates students' likelihood of admission into specific colleges and universities to which they might apply. Our view was that this type of tool could be a means for educators, students, and their families to gain more accurate information about their chances of going to college, which may in turn encourage students to apply to colleges that they may not have been considering before. However, we believed this tool would only be useful if it was a valid and reliable predictor of universities' admissions decisions.

"The results from our statistical models show that students' admissions decisions can be estimated with a high degree of accuracy with a limited set of variables related to students' academic preparedness and controlling for the specific institution they applied to."

The results from our statistical models show that students' admissions decisions can be estimated with a high degree of accuracy with a limited set of variables related to students' academic preparedness and controlling for the specific institution they applied to. The models we developed correctly classified roughly 84% of applications and accurately identified roughly 91% of students who were indeed admitted to college. The models did not perform as well as identifying true negatives; approximately 30% of students who were not admitted would have been told that they would be admitted using this tool. We argue that the

risk of incorrectly telling students they will be admitted is less of a concern than incorrectly telling students that they will not be admitted, as the latter might deter students from applying to institutions they would be admitted to.

Equally important, the results show that controlling for students' demographic backgrounds did not improve the accuracy of the models in any appreciable way, despite these variables being statistically significantly related to the outcome given our large sample size. This finding is important for three reasons. First, the results show that the risk of decreasing model accuracy by excluding demographic controls is minimal. Second, given the ethical concerns of including demographic variables in the interactive tool, which would allow students to see how their race/ethnicity, sex, and economic status influence their likelihood of admissions, the results justify excluding these variables in the interactive tool as well. Third, while debates continue in research, policy, and the courts over affirmative action and how students' demographic backgrounds relate to their odds of admission, our findings suggest that students' demographic characteristics matter little to their likelihood of admission, at least across the full range of public 4-year institutions in Texas.

Most importantly, the statistical models were used to build an interactive tool to demonstrate to students their odds of admission. Given that existing literature has shown many students, and particularly low-income and first-generation students, have limited information about their odds of admission, this tool could be used to help close that information gap. Students who are unsure about their college aspirations or the selectivity of college they aspire to attend may feel encouraged to see first-hand that their academic experiences give them strong chances of admission to a college they are interested in. This predictive admissions tool could therefore be used to increase the selectivity of colleges that low-income and first-generations apply to, and hopefully enroll in, thereby reducing the extent of academic undermatch found consistently in the literature.

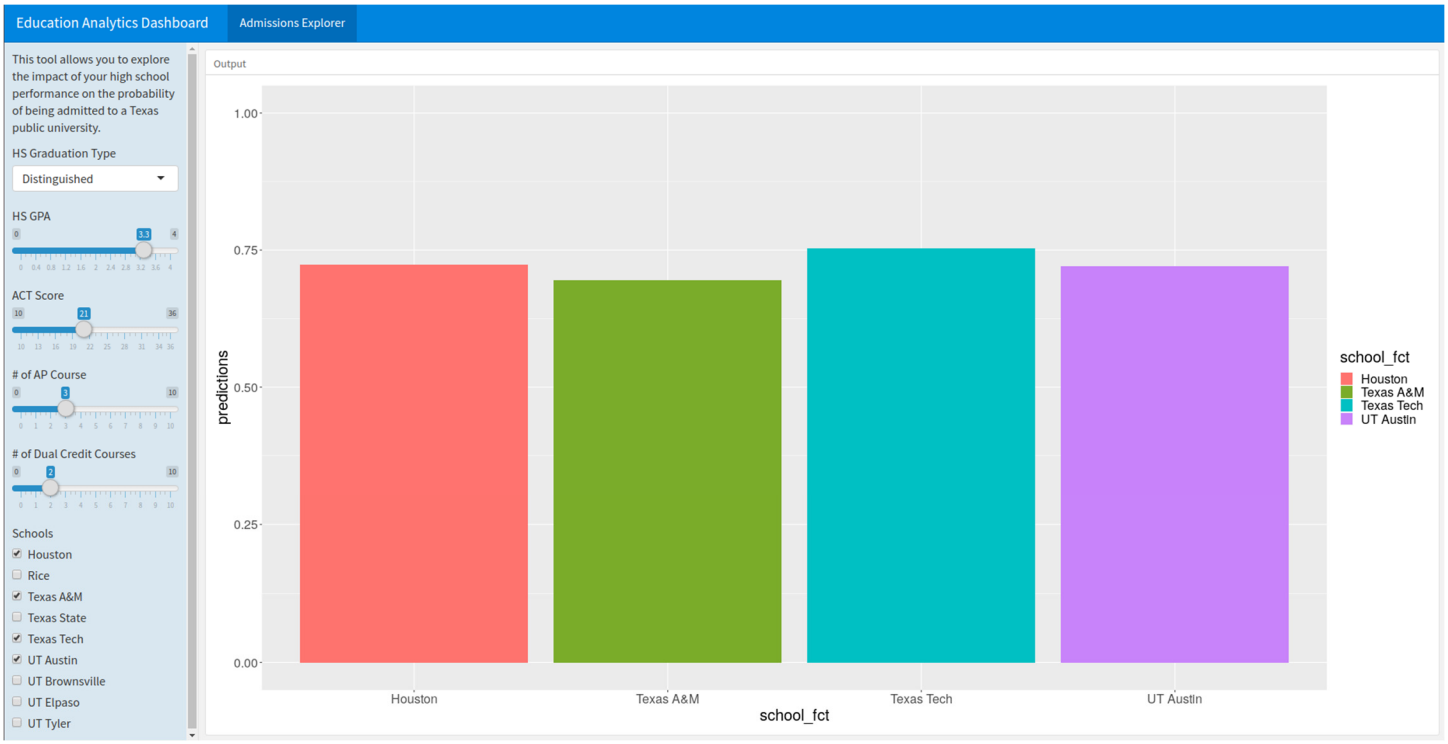
There are a number of ways in which the model we have developed could be broadened. For example, variables such as the highest level of math taken in high school, the number of advanced, dual-credit, and other courses taken by their subject, scores on separate components of standardized tests, and others could easily be added to the model. However, as an initial prototype we opted for the simplest model possible, with positive results. Future research could explore the extent to which more complete and complex models affects their predictive accuracy.

Additionally, further work is need to ensure our model is valid over time. We chose the most recent cohort available at the time we initially began creating the analytic dataset. Replicating this approach with additional cohorts could address a number of intriguing

Table 1.
Results of Logistic Regression Models

	Dependent Variable:							
	admit_fct							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	-28.119***	-28.042***	-28.258***	-28.386***	-28.181***	-28.307**	-28.442***	-28.363***
	-0.212	(0.212)	(0.214)	(0.215)	(0.214)	(0.214)	(0.217)	(0.217)
gpa9_10	2.978***	2.942***	3.000***	2.997***	2.964***	2.961***	3.009***	2.973***
	(0.033)	(0.034)	(0.034)	(0.033)	(0.034)	(0.034)	(0.034)	(0.034)
act	0.491***	0.497***	0.499***	0.499***	0.505***	505***	0.504***	0.509***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
gradtype_fctRegular	0.694	0.691***	0.697***	0.692***	0.694***	0.689***	0.695***	0.693***
	(0.097)	(0.097)	(0.097)	(0.097)	(0.097)	(0.097)	(0.097)	(0.097)
gradtype_fctHonors	0.969***	0.957***	0.963***	0.960***	0.951***	0.948***	0.959***	0.947***
	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)
num_adv_course	0.138***	0.136***	0.137***	0.135***	0.136***	0.133***	0.136***	0.135***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
num_dual_credit	0.064***	0.063***	0.062***	0.063***	0.061***	0.062***	0.061***	0.060***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
male_fct1		-0.164***			-0.165***	-0.163***		-0.164***
		(0.020)			(0.20)	(0.020)		(0.020)
ethnic_group_fctAsian			-0.407***		-0.409***		-0.379***	-0.381***
			(0.046)		(0.046)		(0.047)	(0.047)
ethnic_group_fctBlack			-0.056*		-0.058*		-0.046	-0.049
			(0.030)		(0.030)		(0.030)	(0.030)
ethnic_group_fctOther			-0.431***		-0.431***		-0.403***	-0.404***
			(0.059)		(0.059)		(0.060)	(0.060)
ethnic_group_fctWhite			-0.126***		-0.129***		-0.075***	-0.078***
			(0.027)		(0.027)		(0.029)	(0.029)
econ_dis_fct1				0.162***		0.160***	0.136***	0.134***
				(0.022)		(0.022)	(0.023)	(0.023)
Observations	110,620.00	110,621.00	110,622.00	110,623.00	110,624.00	110,625.00	110,626.00	110,627.00
Log Likelihood	-33,627.06	-33,594.23	-33,565.44	-33,600.41	-33,532.26	-33,568.12	-33,548.41	-33,515.85
Akaike Inf. Crit	67,330.13	67,266.46	67,214.89	67,278.81	67,150.53	67,216.25	67,182.82	67,119.70

Figure 1.
Predictive Admissions Tool Dashboard



research questions, such as whether the model is more or less accurate for other cohorts, whether the relationship between specific academic variables and university admissions has changed over time, and whether demographic variables are more or less impactful during other periods.

References

Adelman, C. (1999). Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment. Washington, DC: National Center for Education Statistics.

Adelman, C. (2006). The toolbox revisited: Paths to degree completion from high school through college. Washington, DC: National Center for Education Statistics.

Alon, S., & Tienda, M. (2005). Assessing the "mismatch" hypothesis: Differences in college graduation rates by institutional selectivity. Sociology of Education, 78(4), 294-315.

An, B. P. (2013). The impact of dual enrollment on college degree attainment: Do low-SES students benefit? Educational Evaluation and Policy Analysis, 35(1), 57-75.

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York, NY: Wiley.

Bound, J., Lovenheim, M. F., & Turner, S. (2010). Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. American Economic Journal: Applied Economics, 2(3), 129-157.

Bourdieu, P. (1988). Homo Academicus. Cambridge, UK: Polity Press.

Bourdieu, P. (1993). The feld of cultural production. Cambridge, UK: Polity Press.

Bowen, W. G., Chingos, M. M., and McPherson, M. S. (2009). Crossing the finish line: Completing college at America's public universities. Princeton, NJ: Princeton University Press.

Brier, G. W. (195). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3.
Cabrera, A. F., & La Nasa, S. M. (2000). Understanding the college-choice process. New Directions for Institutional Research, 107, 5-22.

Cabrera, A. F., & La Nasa, S. M. (2001). On the path to college: Three critical tasks facing America's disadvantaged. Research in Higher Education, 42(2), 119-149.

Carnevale, A. P., & Rose, S. J. (2003). Socioeconomic status, race/ethnicity, and selective college admissions. New York, NY: The Century Foundation.

Cohn, E., Cohn, S., Balch, D. C., & Bradley, J. (2004). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank. Economics of Education Review, 23, 577-586.

College Board. (2016). Higher ed brief: Concordance tables. Retrieved from collegereadiness.collegeboard.org/pdf/higher-ed-brief-sat-concordance.pdf.

Demler, O. V., Pencina, M. J., & D'Agostino, R. B. (2012). Misuse of Delong test to compare AUCs for nested models. Statistics in Medicine, 31(23), 2577-2587.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? American Journal of Sociology, 94(5), 1146-1183.

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. Center for Studies in Higher Education Research & Occasional Paper Series, CSHE.6.07, 1-35.

Giani, M. S. (2015). The postsecondary resource trinity model: Exploring the interaction between socioeconomic, academic, and institutional resources. Research in Higher Education, 56(2), 105-126. DOI 10.1007/s11162-014-9357-4.

Greene, W. H. (2017). Econometric analysis (8th Ed.). New York, NY: Macmillan.

Heyns, B. (1974). Social selection and stratification within schools. American Journal of Sociology, 79(6), 1434-1451.

Hoffman, J. L., & Lowitzki, K. E. (2005). Predicting college success with high school grades and test scores: Limitations for minority students. The Review of Higher Education, 28 (4), 455-474.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd Ed.). New York, NY: John Wiley & Sons, Inc.

Hoxby, C. M., & Avery, C. (2012). The missing "one-offs": The hidden supply of high-achieving, low-income students. National Bureau of Economic Research, Working Paper 18586.

Hoxby, C. M., & Turner, S. (2013). Informing students about their college options: A proposal for broadening the Expanding College Opportunities Project. Washington, DC: Brookings Institution.

Hurtado, S., Inkelas, K. K., Briggs, C., & Rhee, B. (1997). Differences in college access and choice among racial/ethnic groups: Identifying continuing barriers. Research in Higher Education, 38(1), 43-75.

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. Annual Review of Sociology, 29, 417-442.

Kim, D. H. (2004). The effect of financial aid on students' college choice: Differences by racial groups. Research in Higher Education, 45(1), 43-70.

Mattern, K. D., Marini, J. P., & Shaw, E. (2013). Are AP® students more likely to graduate from college on time? New York, NY: College Board.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), FRONTIERS IN ECONOMETRICS (105-142). New York, NY:, Academic Press.

McPherson, M. S., & Shapiro, M. O. (1998). The student aid game: Meeting need and rewarding talent in American higher education. Princeton, NJ: Princeton University Press.

Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595-600.

National Center for Education Statistics. (2016). Digest of Education Statistics, 2015. Washington, DC: Author.
Oakes, J., & Guiton, G. (1995). Matchmaking: The dynamics of high school tracking decisions. American Educational Research Journal, 32(1), 3-33.

Pallais, A., & Turner, S. (2006). Opportunities for low-income students at top colleges and universities: Policy initiatives and the distribution of students. National Tax Journal, 59(2), 357-386.

Roderick, M., Coca, V., & Nagaoka, J. (2012). Potholes on the road to college: High school effects in shaping urban students' participation in college application, four-year college enrollment, and college match. Sociology of Education, 84(3), 178-211. doi:10.1177/003804071141128.

Roderick, M., Nagaoka, J., Coca, V., Moeller, E., Roddie, K., Gilliam, J., & Patton, D. (2008). From high school to the future: Potholes on the fold to college. Chicago, IL: Consortium on Chicago School Research at the University of Chicago.

Rosenbaum, J. E. (1976). Making inequality: The hidden curriculum of high school tracking. New York, NY: John Wiley & Sons.

Sander, R. H., & Taylor, S., Jr. (2012). Mismatch: How affirmative action hurts students it's intended to help, and why universities won't admit it. New York, NY: Basic Books.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417-453.

Smith, J., Pender, M., & Howell, J. (2012). The full extent of student-college academic undermatch. Economics of Education Review, 32, 247-261.

Speroni, C. (2011). Determinants of students' success: The role of advanced placement and dual enrollment programs. New York, NY: National Center for Postsecondary Research, Teachers College, Columbia University. Retrieved from postsecondaryresearch.org/publications/19811_Speroni _AP_DE_paper_110311_FINAL.pdf.

Terenzini, P. T., Cabrera, A. F., & Bernal, E. M. (2001). Swimming against the tide: The poor in American higher education. College Board Research Report No. 2001-1. New York, NY: The College Board.

Thernstrom, S., & Thernstrom, A. (1997). America in Black and White: One nation indivisible. New York, NY: Simon & Schuster.

White, K. (1982). The relation between socioeconomic status and academic achievement. Psychological Bulletin, 91, 461–481.

Zwick, R., & Sclar, J. G. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. American Educational Research Journal, 42(3), 439-464.